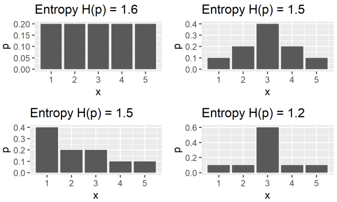
Introduction to Machine Learning

Joint Entropy and Mutual Information



Learning goals

- Know the joint entropy
- Know conditional entropy as remaining uncertainty
- Know mutual information as the amount of information of an RV obtained by another

JOINT ENTROPY

- The **joint entropy** of two discrete random variables X and Y with a joint distribution $p(x, y)$ is:

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log(p(x, y)),$$

  which can also be expressed as

$$H(X, Y) = -\mathbb{E}\left[\log(p(X, Y))\right].$$

- For continuous random variables X and Y with joint density $p(x, y)$, the differential joint entropy is:

$$h(X, Y) = -\int_{\mathcal{X}, \mathcal{Y}} p(x, y) \log p(x, y) dxdy$$

For the rest of the section we will stick to the discrete case. Pretty much everything we show and discuss works in a completely analogous manner for the continuous case - if you change sums to integrals.

CONDITIONAL ENTROPY

- The **conditional entropy** $H(Y|X)$ quantifies the uncertainty of Y that remains if the outcome of X is given.
- $H(Y|X)$ is defined as the expected value of the entropies of the conditional distributions, averaged over the conditioning RV.
- If $(X, Y) \sim p(x, y)$, the conditional entropy $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) = \mathbb{E}_X[H(Y|X=x)] &= \sum_{x \in \mathcal{X}} p(x) H(Y|X=x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x) \\ &= -\mathbb{E}\left[\log p(Y|X)\right]. \end{aligned}$$

- For the continuous case with density f we have

$$h(Y|X) = -\int f(x,y) \log f(x|y) dxdy.$$

CHAIN RULE FOR ENTROPY

The **chain rule for entropy** is analogous to the chain rule for probability and, in fact, derives directly from it.

$$H(X, Y) = H(X) + H(Y|X)$$

Proof:
$$\begin{aligned}
H(X, Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log p(x, y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log p(x)p(y|x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x) \log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

n-Variable version:

$$H\left(X_1, X_2, \ldots, X_n\right) = \sum_{i=1}^{n} H\left(X_i | X_{i-1}, \ldots, X_1\right).$$

JOINT AND CONDITIONAL ENTROPY

The following relations hold:

$$\begin{aligned} H(X, X) &= H(X) \\ H(X|X) &= 0 \\ H(X, Y|Z) &= H(X|Z) + H(Y|X, Z) \end{aligned}$$

Which can all be trivially derived from the previous considerations.

Furthermore, if $H(X|Y) = 0$, then X is a function of Y, so for all y with $p(y) > 0$, there is only one x with $p(x, y) > 0$. Proof is not hard, but also not completely trivial.

MUTUAL INFORMATION

- The MI describes the amount of information about one random variable obtained through the other one or how different the joint distribution is from pure independence.
- Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The MI $I(X; Y)$ is the Kullback-Leibler Divergence between the joint distribution and the product distribution $p(x)p(y)$:

$$\begin{aligned} I(X;Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D_{KL}(p(x,y) \| p(x)p(y)) \\ &= \mathbb{E}_{p(x,y)} \left[\log \frac{p(X,Y)}{p(X)p(Y)} \right]. \end{aligned}$$

- For two continuous random variables with joint density $f(x, y)$:

$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy.$$

MUTUAL INFORMATION

We can rewrite the definition of mutual information $I(X;Y)$ as

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \\
&= -\sum_{x} p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x|y) \right) \\
&= H(X) - H(X|Y).
\end{aligned}
$$

Thus, mutual information $I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y.

MUTUAL INFORMATION

The following relations hold:

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
I(X;Y) &= H(Y) - H(Y|X) \\
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
I(X;Y) &= I(Y;X) \\
I(X;X) &= H(X)
\end{aligned}
$$

All of the above are trivial to prove.

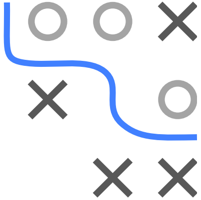
MUTUAL INFORMATION - EXAMPLE

Let X, Y have the following joint distribution:

	X_1	X_2	X_3	X_4
Y_1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
Y_2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
Y_3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
Y_4	$\frac{1}{4}$	0	0	0

The marginal distribution of X is $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ and the marginal distribution of Y is $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, and hence $H(X) = \frac{7}{4}$ bits and $H(Y) = 2$ bits.

MUTUAL INFORMATION - EXAMPLE



The conditional entropy $H(X|Y)$ is given by:

$$
\begin{aligned}
H(X|Y) &= \sum_{i=1}^{4} p(Y=i) H(X|Y=i) \\
&= \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) \\
&+ \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4} H(1, 0, 0, 0) \\
&= \frac{1}{4} \cdot \frac{7}{4} + \frac{1}{4} \cdot \frac{7}{4} + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 0 \\
&= \frac{11}{8} \text{ bits.}
\end{aligned}
$$

Similarly, $H(Y|X) = \frac{13}{8}$ bits and $H(X, Y) = \frac{27}{8}$ bits.

MUTUAL INFORMATION - COROLLARIES

Non-negativity of mutual information: For any two random variables, X, Y, $I(X; Y) \geq 0$, with equality if and only if X and Y are independent.

Proof: $I(X; Y) = D_{KL}(p(x, y) \| p(x)p(y)) \geq 0$, with equality if and only if $p(x, y) = p(x)p(y)$ (i.e., X and Y are independent).

Conditioning reduces entropy (information can't hurt):

$$H(X|Y) \leq H(X),$$

with equality if and only if X and Y are independent.

Proof: $0 \leq I(X; Y) = H(X) - H(X|Y)$
Intuitively, the theorem says that knowing another random variable Y can only reduce the uncertainty in X. Note that this is true only on the average.

MUTUAL INFORMATION - COROLLARIES

Independence bound on entropy: Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i),$$

with equality if and only if the X_i are independent.

Proof: With the chain rule for entropies,

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \quad \leq \sum_{i=1}^{n} H(X_i),$$

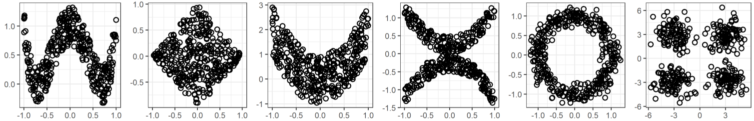
where the inequality follows directly from above. We have equality if and only if X_i is independent of $X_{i-1}, \ldots, X_1$ for all i (i.e., if and only if the X_i 's are independent).

MUTUAL INFORMATION PROPERTIES

- MI is a measure of the amount of "dependence" between variables. It is zero if and only if the variables are independent.
- On the other hand, if one of the variables is a deterministic function of the other, the mutual information is maximal, i.e. entropy of the first.
- Unlike (Pearson) correlation, mutual information is not limited to real-valued random variables.
- Mutual information can be used to perform **feature selection**. Quite simply, each variable X_i is rated according to $I(X_i; Y)$, this is sometime called information gain.
- The same principle can also used in decision trees to select a feature to split on. Splitting on MI/IG is then equivalent to risk reduction with log-loss.

MUTUAL INFORMATION VS. CORRELATION

- If two variables are independent, their correlation is 0.
- However, the reverse is not necessarily true. It is possible for two dependent variables to have 0 correlation because correlation only measures linear dependence.



- The figure above shows various scatterplots where, in each case, the correlation is 0 even though the two variables are strongly dependent, and MI is large.
- Mutual information can therefore be seen as a more general measure of dependence between variables than correlation.

MUTUAL INFORMATION - EXAMPLE

Let X, Y be two correlated Gaussian random variables.
$(X, Y) \sim \mathcal{N}(0, K)$ with correlation ρ and covariance matrix K:

$$K = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

Then $h(X) = h(Y) = \frac{1}{2} \log\left((2\pi e)\sigma^2\right)$, and
$h(X, Y) = \frac{1}{2} \log\left((2\pi e)^2 |K|\right) = \frac{1}{2} \log\left((2\pi e)^2 \sigma^4 (1 - \rho^2)\right)$, and thus

$$I(X; Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2} \log(1 - \rho^2).$$

For $\rho = 0$, X and Y are independent and $I(X; Y) = 0$.
For $\rho = \pm 1$, X and Y are perfectly correlated and $I(X; Y) \to \infty$.